(1)

$A_{in} = 5000 \times 100$

$W_1 = 128 \times 5000$

$b_1 = 128 \times 1$

$A_1 = 128 \times 100$

$W_{out} = 5000 \times 128$

$b_{out} = 5000 \times 1$

$A_{out} = 5000 \times 100$

(2) $A_1 = relu(Z_1)$

$$relu(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x <= 0 \end{cases}$$

$$= \begin{bmatrix} 12 & 4 \\ 2 & 16 \\ 13 & 2 \\ 0 & 1 \end{bmatrix}$$

$Z_{out} = W_{out} A_1 + b_{out} =$

$$\begin{bmatrix} 3 & -1 & 2 & -4 \\ 1 & -5 & -1 & 3 \end{bmatrix} \begin{bmatrix} 12 & 4 \\ 2 & 16 \\ 13 & 2 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 4 \\ -5 \end{bmatrix}$$

broadcasted!

$$= \begin{bmatrix} 64 & 0 \\ -16 & -80 \end{bmatrix}$$

$f_{out} = relu$

$A_{out} = f_{out}(Z_{out})$

$$= \begin{bmatrix} 64 & 0 \\ 0 & 0 \end{bmatrix}$$

(2)

$$Z_1 = W_1 A_{in} + b_1 = \begin{bmatrix} 3 & 2 & -1 & 0 & -3 & 2 \\ 4 & -1 & -2 & 3 & 5 & -6 \\ 2 & -1 & 3 & 4 & -3 & 1 \\ 6 & -4 & 2 & -5 & 1 & 2 \end{bmatrix}$$

$$\times \begin{bmatrix} 1 & 2 \\ 4 & 5 \\ 1 & 3 \\ 3 & 1 \\ 1 & 4 \\ 2 & 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \end{bmatrix} \text{ – broadcasted!}$$

$$= \begin{bmatrix} 12 & 4 \\ 2 & 16 \\ 13 & 2 \\ -16 & 1 \end{bmatrix}$$

3a

$$\frac{\partial Loss}{\partial W_{out}} = \frac{\partial Loss}{\partial A_{out}} \circ \frac{\partial A_{out}}{\partial Z_{out}} \circ \frac{\partial Z_{out}}{\partial W_{out}}$$

$$\frac{\partial Loss}{\partial b_{out}} = \frac{\partial Loss}{\partial A_{out}} \circ \frac{\partial A_{out}}{\partial Z_{out}} \circ \frac{\partial Z_{out}}{\partial b_{out}}$$

$$\frac{\partial Loss}{\partial A_1} = \frac{\partial Loss}{\partial A_{out}} \circ \frac{\partial A_{out}}{\partial Z_{out}} \circ \frac{\partial Z_{out}}{\partial A_1}$$

$$\frac{\partial Loss}{\partial W_1} = \frac{\partial Loss}{\partial A_{out}} \circ \frac{\partial A_{out}}{\partial Z_{out}} \circ \frac{\partial Z_{out}}{\partial A_1} \circ \frac{\partial A_1}{\partial Z_1} \circ \frac{\partial Z_1}{\partial W_1}$$

$$\frac{\partial Loss}{\partial b_1} = \frac{\partial Loss}{\partial A_{out}} \circ \frac{\partial A_{out}}{\partial Z_{out}} \circ \frac{\partial Z_{out}}{\partial A_1} \circ \frac{\partial A_1}{\partial Z_1} \circ \frac{\partial Z_1}{\partial b_1}$$

(3b)

$$Z_1 = \begin{bmatrix} 12 & 4 \\ 2 & 16 \\ 13 & 2 \\ -16 & 1 \end{bmatrix} \qquad A_1 = \begin{bmatrix} 12 & 4 \\ 2 & 16 \\ 13 & 2 \\ 0 & 1 \end{bmatrix}$$

$$Z_{out} = \begin{bmatrix} 64 & 0 \\ -16 & -80 \end{bmatrix} \qquad A_{out} = \begin{bmatrix} 64 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\frac{\partial Loss}{\partial W_{out}} = \begin{bmatrix} 3 & 5 \\ 2 & -4 \end{bmatrix} \circ relu\_der(Z_{out}) \circ A_1^T$$

mask

$$= \begin{bmatrix} 3 & 5 \\ 2 & -4 \end{bmatrix} \circ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \circ \begin{bmatrix} 12 & 2 & 13 & 0 \\ 4 & 16 & 2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix} \circ \begin{bmatrix} 12 & 2 & 13 & 0 \\ 4 & 16 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 36 & 6 & 39 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

mask

$$\frac{\partial Loss}{\partial b_{out}} = \begin{bmatrix} 3 & 5 \\ 2 & -4 \end{bmatrix} \circ relu\_der(Z_{out}) \circ \begin{bmatrix} 1 & 1 \end{bmatrix}^T$$

$$= \begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix} \circ \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$$

(3b) $\frac{\partial Loss}{\partial W_1} = \begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix} \circ W_{out} \circ \text{relu\_der}(Z_1) \circ A_{in}^T$ (mask)

$= \begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix} \circ \begin{bmatrix} 3 & -1 & 2 & -4 \\ 1 & -5 & -1 & 3 \end{bmatrix} \circ \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \circ A_{in}^T$

$= \begin{bmatrix} 9 & -3 & 6 & -12 \\ 0 & 0 & 0 & 0 \end{bmatrix}^T \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} A_{in}^T = \begin{bmatrix} 9 & 0 \\ -3 & 0 \\ 6 & 0 \\ -12 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} A_{in}^T$ (mask)

$= \begin{bmatrix} 9 & 0 \\ -3 & 0 \\ 6 & 0 \\ 0 & 0 \end{bmatrix} \circ A_{in}^T = \begin{bmatrix} 9 & 36 & 9 & 27 & 9 & 18 \\ -3 & -12 & -3 & -9 & -3 & -6 \\ 6 & 24 & 6 & 18 & 6 & 12 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$

how we get the last element

$$\begin{cases} Z_{out} = W_{out} A_1 + b_{out} & [b_{out}\ b_{out}] \\ \frac{\partial Z}{\partial b} = [1\ \ 1] \end{cases}$$

same as in $\partial Loss / \partial W_1$

(3b)

$$\frac{\partial Loss}{\partial b_1} = \overbrace{\frac{\partial Loss}{\partial A_{out}} \cdot \frac{\partial A_{out}}{\partial Z_{out}} \cdot \frac{\partial Z_{out}}{\partial A_1} \cdot \frac{\partial A_1}{\partial Z_1}} \cdot \frac{\partial Z_1}{\partial b_1}$$

$$\Rightarrow \begin{bmatrix} 9 & 0 \\ -3 & 0 \\ 6 & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 9 \\ -3 \\ 6 \\ 0 \end{bmatrix}$$

(3c) Updated values

learning rate: 0,01

$$W_{out\_updated} = W_{out} - 0{,}01 \cdot \frac{\partial Loss}{\partial W_{out}} =$$

$$\begin{bmatrix} 3 & -1 & 2 & -4 \\ 1 & -5 & -1 & 3 \end{bmatrix} - \begin{bmatrix} 0{,}35 & 0{,}06 & 0{,}39 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 2{,}64 & -1{,}06 & 1{,}61 & -4 \\ 1 & -5 & -1 & 3 \end{bmatrix}$$

$$b_{out\_updated} = b_{out} - 0{,}01 \cdot \frac{\partial Loss}{\partial b_{out}} =$$

$$\begin{bmatrix} 4 \\ -5 \end{bmatrix} - \begin{bmatrix} 0{,}03 \\ 0 \end{bmatrix} = \begin{bmatrix} 3{,}97 \\ -5 \end{bmatrix}$$

(3c) $W_1\text{-updated} = W_1 - 0{,}01 \cdot \frac{\partial Loss}{\partial W_1}$

$$= \begin{bmatrix} 3 & 2 & -1 & 0 & -3 & 2 \\ 4 & -1 & -2 & 3 & 5 & -6 \\ 2 & -1 & 3 & 4 & -3 & 1 \\ 6 & -4 & 2 & -5 & 1 & 2 \end{bmatrix}$$

$$- \begin{bmatrix} 0{,}09 & 0{,}36 & 0{,}09 & 0{,}27 & 0{,}09 & 0{,}18 \\ -0{,}03 & -0{,}12 & -0{,}03 & -0{,}09 & -0{,}03 & -0{,}06 \\ 0{,}06 & 0{,}24 & 0{,}06 & 0{,}18 & 0{,}06 & 0{,}12 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 2{,}91 & 1{,}64 & -1{,}09 & -0{,}27 & -3{,}09 & 1{,}82 \\ 4{,}03 & -0{,}88 & -1{,}97 & 3{,}09 & 5{,}03 & -5{,}94 \\ 1{,}94 & -1{,}24 & 2{,}94 & 3{,}82 & -3{,}06 & 0{,}88 \\ 6 & -4 & 2 & -5 & 1 & 2 \end{bmatrix}$$

$$b_1\text{-updated} = b_1 - 0{,}01 \cdot \frac{\partial Loss}{\partial b_1}$$

$$= \begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 0{,}09 \\ -0{,}03 \\ 0{,}06 \\ 0 \end{bmatrix} = \begin{bmatrix} 0{,}91 \\ 2{,}03 \\ 0{,}94 \\ 2 \end{bmatrix}$$

Question 4. The GloVe and word2vec libraries provide pre-trained word embeddings for dimensions from 50 - 300. In an NLP application task what are the trade-offs of using word embeddings of smaller vs larger dimensions? (5 points)

Word2Vec is used to project words in vector space as word embeddings. Glove (Global Vectors) is, in effect, an extension of Word2Vec.

The trade-off of using embeddings of different dimensions, is between a number of things. Firstly, the more dimensions we use, the more computation power (and time) is required to train the model. Secondly, higher dimensionality typically gives better quality of the embeddings. One interpretation is that at 50, quality is typically low, whereas over 300 it usually doesn't increase in a significant manner (hence the interval mentioned in the assignment).

The dimensions of word2vec don't have a clear meaning and as such it is hard to say what features are lost or gained as the dimensions change. Overall, less dimensions mean more general and less accurate embeddings, whereas more dimensions give more specific and accurate embeddings. While accuracy might seem like an elusive concept in this context, the developer of word2vec, Thomas Mikolov, used about 9,000 semantic and 10,000 syntactic relations to produce a benchmark test of the accuracy of the model.